ISSN 2311-5149

Отримано: 21 листопада 2021 р.

Прорецензовано: 30 листопада 2021 р.

Прийнято до друку: 05 грудня 2021 р.

e-mail: oleksandr.novoseletskyy@oa.edu.ua

DOI: 10.25264/2311-5149-2021-23(51)-118-123

Новоселецький О. М., Гончарова В. О. Ідентифікації потенційного споживача продукції ринку електронної комерції методом градієнтного бустінгу. *Наукові записки Національного університету «Острозька академія». Серія «Економіка» : науковий журнал.* Острог : Вид-во НаУОА, грудень 2021. № 23(51). С. 118–123.

УДК: 330.4; 519.86

JEL-класифікація: C 45, C 50

ORCID-ідентифікатор: orcid.org/ 0000-0003-3757-0552

ORCID-ідентифікатор: orcid.org/0000-0002-1384-1007

Новоселецький Олександр Миколайович,
кандидат економічних наук, доцент кафедри економіко-математичного моделювання та ІТ Національного університету «Острозька академія»
Гончарова Вікторія Олександрівна,
студентка Національного університету «Острозька академія»

ІДЕНТИФІКАЦІЇ ПОТЕНЦІЙНОГО СПОЖИВАЧА ПРОДУКЦІЇ РИНКУ ЕЛЕКТРОННОЇ КОМЕРЦІЇ МЕТОДОМ ГРАДІЄНТНОГО БУСТІНГУ

Послуги електронної комерції є однією зі сфер, де щодня збираються нові дані. Тому видається необхідним використовувати методи аналізу даних у цій сфері. Взаємодії користувачів з платформами електронної комерції складають складні моделі поведінки, які, якщо їх проаналізувати, можуть дати суб'єктам господарювання можливість зрозуміти потреби споживачів. У статті проведено результати дослідження щодо ідентифікації споживача продукції на ринку електронної комерції за допомогою методу градієнтного бустінгу.

Ключові слова: *поведінка споживачів, градієнтний бустінг, прогнозування покупок.*

Новоселецкий Александр Николаевич,
кандидат экономических наук, доцент кафедры экономико-математического моделирования и информационных технологий Национального университета «Острожская академия»
Гончарова Виктория Александровна,
студентка Национального университета «Острожская академия»

ИДЕНТИФИКАЦИИ ПОТЕНЦИОННОГО ПОТРЕБИТЕЛЯ ПРОДУКЦИИ РЫНКА ЭЛЕКТРОННОЙ КОММЕРЦИИ МЕТОДОМ ГРАДИЕНТНОГО БУСТИНГА

Услуги электронной коммерции являются одной из сфер, где каждый день собираются новые данные. Поэтому представляется нужным употреблять способы анализа данных в данной сфере. Взаимодействия пользователей с платформами электронной коммерции составляют сложные модели поведения, которые, если их проанализировать, могут дать предприятиям возможность понять потребности потребителей. В статье рассмотрен пример применения алгоритма градиентного бустинга для идентификации потребителя на рынке электронной коммерции.

Ключевые слова: *поведение потребителей, градиентный бустинг, прогнозирование покупок.*

Oleksander Novoseletskyy,
Candidate of Economic Sciences, Associate Professor of the Economics of Mathematical Modeling and IT department, The National University of Ostroh Academy
Victoriia Honcharova,
student, The National University of Ostroh Academy

IDENTIFICATION OF A POTENTIAL CONSUMER OF E-COMMERCE MARKET PRODUCTS BY GRADIENT BUSTING METHOD

E-commerce is an integral part of the developed economy in the country. Small and large businesses can sell their products or services online, meeting the needs of consumers anywhere and anytime. The development of e-commerce is impossible without the knowledge of consumer behavior. Data has become one of the world's most valuable resources due to the rapid digital transformation of global industries. Collecting customer data has become a top priority for businesses. As more and more advanced technologies are developed to collect and analyze customer data, more companies are able to contextualize, retrieve and monetize information from them. Knowing why people buy, companies can grow more effectively in e-commerce and do so strategically, knowing what next steps to take. From consumer behavior to predictive analytics, companies regularly collect, store and analyze large amounts of quantitative and qualitative data about their consumer

base on a daily basis. E-commerce services are one of the areas where new data is collected every day. Therefore, it seems necessary to use data analysis methods in this area. User interactions with e-commerce platforms are complex patterns of behavior that, if analyzed, can enable businesses to understand consumer needs. Consumer buying behavior is influenced by many factors, and different consumer demands lead to large differences in consumer buying behavior. To predict consumer buying behavior it is necessary to determine the hidden characteristics of data in the array of information left by users on the e-commerce platform, and then to determine the desire of future users to buy on the e-commerce platform. The article considers an example of application of the gradient boosting algorithm for consumer identification in the e-commerce market.

Keywords: *consumer behavior, gradient boosting, shopping forecasting.*

Постановка проблеми. Останніми роками ринок електронної комерції розвивається все швидше. У складному та хаотичному ринковому середовищі ринок електронної комерції стикається з можливостями та викликами. Користувачі залишають значну кількість даних на платформах електронної комерції, але лише невелика кількість даних перетворюється на купівельну поведінку. Велике значення в дослідженні електронної комерції має використання алгоритмів машинного навчання. У цьому дослідженні впроваджується алгоритм градієнтного бустінгу (XGBoost) для ідентифікації потенційного споживача продукції ринку електронної комерції. За допомогою інструментів R досліджуються та аналізуються характеристики даних про наміри придбати споживачів платформи електронної комерції, що може надати підприємствам новий метод аналізу та прогнозування поведінки споживачів.

Аналіз останніх досліджень та публікацій. Серед перших досліджень в сфері електронної комерції можна виділити роботу Д. Козьє «Електронна комерція», де автор розглянув сучасні електронні бізнес-технології, проаналізував досвід компаній, діяльність яких відбувається в сфері e-commerce. Також можна виокремити роботу Д. Еймора «Електронний бізнес. Еволюція і/чи революція. Життя та бізнес в епоху Інтернету», навчальний посібник І. Т. Балабанова «Електронна комерція». О. В. Мельник вивчав електронну комерцію як складову частину цифрової економіки, С. І. Ляпунов досліджував глобальний бізнес та інформаційні технології, І. Н. Успенський написав «Енциклопедію інтернет-бізнесу». Питаннями математичного моделювання поведінки споживачів займаються як вітчизняні, так і зарубіжні дослідники: Б. Ліпштейн виділяв особливий вплив реклами на вибір споживача, його лояльність. С. Патель, А. Шлігер у імовірнісній моделі враховували психологічні особливості поведінки споживача при купівлі товару. До складу моделі вони включили ефекти, які описують ці деталі.

Мета статті: розробити економетричну модель ідентифікації потенційного споживача на ринку електронної комерції за допомогою алгоритму градієнтного бустінгу.

Виклад основного матеріалу дослідження. Електронна комерція є однією з найбільш популярних бізнес-моделей у всьому світі. Люди можуть замовити товари і отримати їх за кілька днів або, можливо, годин, не відвідуючи магазини.

Важливим поняттям електронної комерції є поведінка споживачів в інтернеті. Поведінка споживачів в інтернеті – це процес того, як споживачі приймають рішення про покупку продуктів в електронній комерції [1].

Послуги електронної комерції є однією зі сфер, де щодня збираються нові дані. Тому видається необхідним використовувати методи аналізу даних у цій сфері. У цій статті наведений приклад розробки моделі ідентифікації споживача на ринку електронної комерції методом градієнтного бустінгу.

XGBoost – це алгоритм, який реалізує ефективну класифікацію під структурою підвищення градієнта. Він покращує GBM, яка володіє характеристиками високої ефективності, гнучкості та портативності, а також може створити дерево рішень із прискореним градієнтом. Основною ідеєю GBM є ідея градієнтного спуску, при якому кожне згенероване дерево базується на попередньому результаті для мінімізації цільової функції [2].

Припускаючи, що даний набір даних – D, кількість вибірок – n, а кількість власних значень – m, $D = \{(X_i, y_i)\}$, $(|D| = n, X_i \in R^m, y_i \in R)$. Основна структура моделі XGBoost генерується моделлю додавання K-моделей дерева, кожне дерево відповідає залишкам попереднього дерева. Інтегровану модель дерева можна виразити так:

$$y^\wedge_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F$$

$$F = \{f(x) = w_{q(x)}\} (q: R^m \to T, w \in R)$$

де F представляє простір моделі дерева регресії, x_i представляє власні вектори даних, T представляє кількість листових вузлів у дереві, а f_k представляє незалежну структуру дерева q і вагу листка w.

Оскільки ми будемо використовувати алгоритми машинного в R, нам не доведеться самостійно рахувати всі функції. Для здійснення класифікації за допомогою градієнтного бустінгу використаємо бібліотеку Xgboost.

Для ідентифікації споживача електронної комерції використаємо відкриту базу даних Kaggle [3]. В ній міститься інформація про споживачів електронної комерції, які переглядали товари на сайті певної компанії та здійснили або не здійснили покупку. Компанія займається продажем різноманітних товарів, зокрема одягу, речей для дому та ін.

Використаємо для побудови моделі такі змінні:

- Credit Score – показник здатності особи повернути позичену суму;
- Geography – країни, в яких знаходяться користувачі (Франція, Німеччина, Іспанія);
- Gender – стать користувача;
- Age – вік;
- Tenure – термін перебування;
- Balance – баланс;
- NumOfProducts – кількість продуктів;
- HasCrCard – чи має користувач кредитну картку;
- IsActiveMember – чи є активним покупцем;
- EstimatedSalary – орієнтовна заробітна плата;
- Purchase – чи була здійснена покупка (результуюча змінна).

Спочатку нам необхідно здійснити попередній аналіз даних. Дослідницький аналіз даних є одним із важливих процесів виконання початкових досліджень. Основна ідея полягає в тому, щоб виявити закономірності, аномалії, перевірити гіпотези та перевірити припущення за допомогою підсумкової статистики та графічного зображення.

Для перегляду даних скористаємось функцією str. Як можемо бачити нижче, датасет містить 10 тис. спостережень та різні види змінних. Серед змінних є категоріальні, біноміальні та числові значення.

```
str(dataset)

## 'data.frame':    10000 obs. of  11 variables:
##  $ CreditScore    : int  619 608 502 699 850 645 822 376 501 684 ...
##  $ Geography      : chr  "France" "Spain" "France" "France" ...
##  $ Gender         : chr  "Female" "Female" "Female" "Female" ...
##  $ Age            : int  42 41 42 39 43 44 50 29 44 27 ...
##  $ Tenure         : int  2 1 8 1 2 8 7 4 4 2 ...
##  $ Balance        : num  0 83808 159661 0 125511 ...
##  $ NumOfProducts  : int  1 1 3 2 1 2 2 4 2 1 ...
##  $ HasCrCard      : int  1 0 1 0 1 1 1 1 0 1 ...
##  $ IsActiveMember : int  1 1 0 0 1 0 1 0 1 1 ...
##  $ EstimatedSalary: num  101349 112543 113932 93827 79084 ...
##  $ Purchase       : int  1 0 1 0 0 1 0 1 0 0 ...
```

Рис. 1. Перегляд даних

На рис. 2 представлений частотний аналіз категоріальних змінних. Проаналізувавши категоріальні змінні, можемо зробити висновок, що фактор Geography містить 3 унікальних значення. 50,14 % – клієнти з Франції, 25,09 % – клієнти з Німеччини та 24,77 % – клієнти з Іспанії. Фактор Gender містить 2 унікальні значення. Серед клієнтів 54,57 % – чоловіки, 45,43 % – жінки.

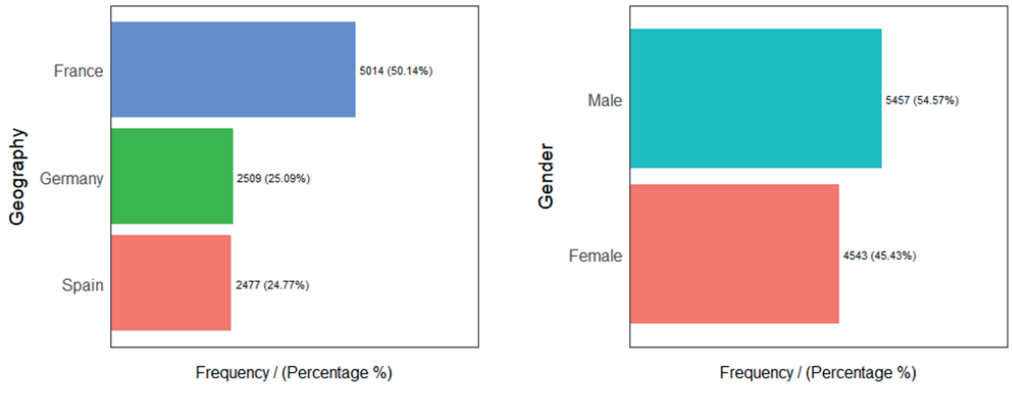

Рис. 2. Візуалізація частотного аналізу для змінних Geograohy та Gender

Частотний аналіз числових змінних можна побачити на рис. 3. За допомогою цих графіків ми отримуємо інформацію про те, які значення приймають змінні, кількість спостережень, які приймають ті чи інші значення.

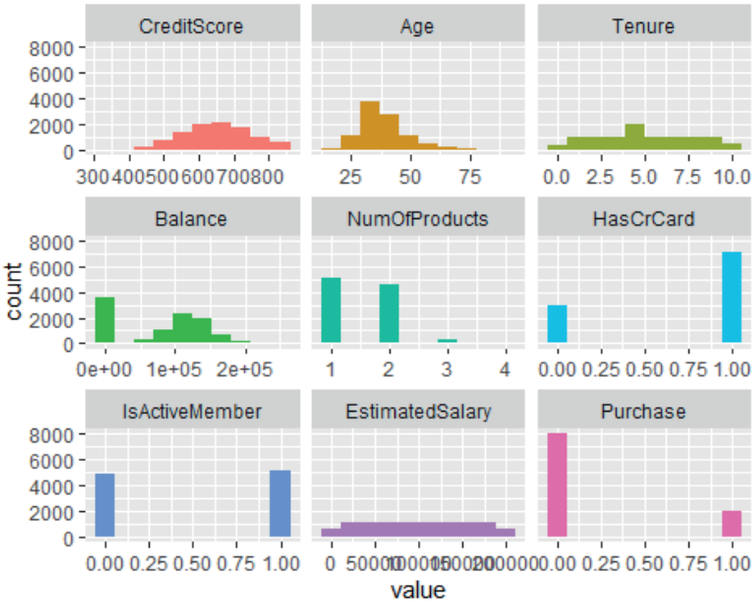

Рис. 3. Візуалізація частотного аналізу числових факторів

Перед тим, як будувати модель, нам необхідно поділити вибірку на тренувальну та тестову. Навчальний набір даних – підмножина для навчання моделі, тестовий набір – підмножина для перевірки навченої моделі.

Тестовий набір має відповідати таким двом умовам:

• Досить великий, щоб отримувати статистично значущі результати.

• Представляє набір даних в цілому. Іншими словами: не вибирайте тестовий набір з характеристиками, відмінними від навчального.

Мета такого поділу – створити модель, яка добре узагальнює нові дані. Наш тестовий набір служить проксі для нових даних [4].

Ми будемо використовувати метод градієнтного бустінгу для ідентифікації потенційного споживача електронної комерції. За результуючу змінну ми взяли біноміальний фактор Purchace, яка показує чи була здійснена покупка на сайті певної компанії.

Особливість цього методу класифікації полягає в тому, що ми не маємо математичної інтерпретації моделі, але можемо оцінити, наскільки добре вона класифікує нові дані, які не входили до процесу навчання моделі.

Confusion Matrix – це таблиця, яка часто використовується для опису ефективності моделі класифікації (або «класифікатора») для набору тестових даних, для яких відомі справжні значення. [5].

Використаємо Confusion Matrix, щоб перевірити, скільки клієнтів наша модель ідентифікувала правильно. За результатами Confusion Matrix в табл. 1 модель правильно ідентифікувала 1541 з 1752 клієнтів, які нічого не придбали на сайті, та 196 з 248 клієнтів, які здійснили покупку. Це свідчить про досить хороші результати. За допомогою Confusion Matrix можемо також визначити точність нашої моделі, яка дорівнює 88,1 %

Таблиця 1

Результати ідентифікації клієнтів моделлю за Confusion Matrix

	Клієнти, які здійснили покупку (фактично)	**Клієнти, які не здійснили покупку (фактично)**
Клієнти, які здійснили покупку (за моделлю)	1541	52
Клієнти, які не здійснили покупку (за моделлю)	211	196

Ще однією важливою оцінкою моделі є крива ROC. Крива AUC-ROC є вимірюванням продуктивності для проблем класифікації при різних порогових налаштуваннях. ROC – це крива ймовірності, а AUC – ступінь або міра відокремленості. Він говорить про те, наскільки модель здатна розрізняти класи. Чим вище AUC, тим краще модель прогнозує 0 класів як 0 і 1 класів як 1. За аналогією, чим вище AUC, тим краще модель розрізняє клієнтів, які здійснили або не здійснили покупку на сайті.

Значення ефективності кривої ROC для нашої моделі дорівнює 0,72. Крива ROC має вигляд:

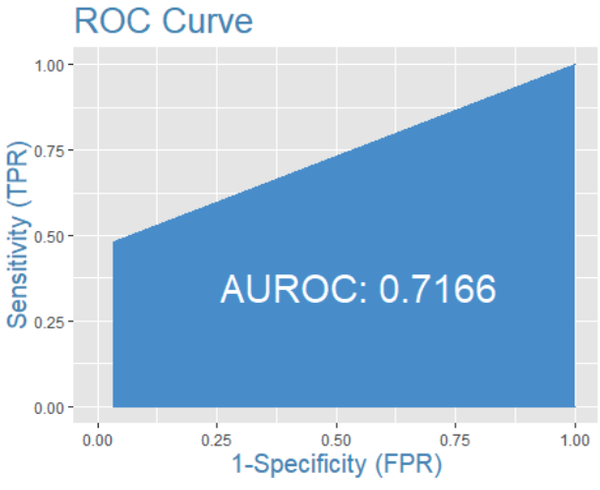

Рис. 3.8. Крива ROC для моделі ідентифікації потенційного споживача на ринку електронної комерції

Тест на придатність Колмогорова-Смірнова порівнює ваші дані з відомим розподілом і дає вам знати, чи мають вони однаковий розподіл. Хоча тест є непараметричним, він не передбачає якогось конкретного основного розподілу – він зазвичай використовується як тест на нормальність, щоб перевірити, чи ваші дані нормально розподілені. Він також використовується для перевірки припущення нормальності в аналізі дисперсій.

Точніше, тест порівнює відомий гіпотетичний розподіл ймовірностей (наприклад, нормальний розподіл) з розподілом, створеним даними – емпіричною функцією розподілу [6].

Чим вищою є статистика Колмогорова-Смірнова, тим ефективнішою є модель при класифікації. В нашому випадку статистика має значення 0,42.

Скористаємось KS Plot для оцінки моделі. KS Plot показує, що буде у випадку, коли ми використовуємо та не використовуємо модель. Чим вище крива моделі, тим кращою вона є.

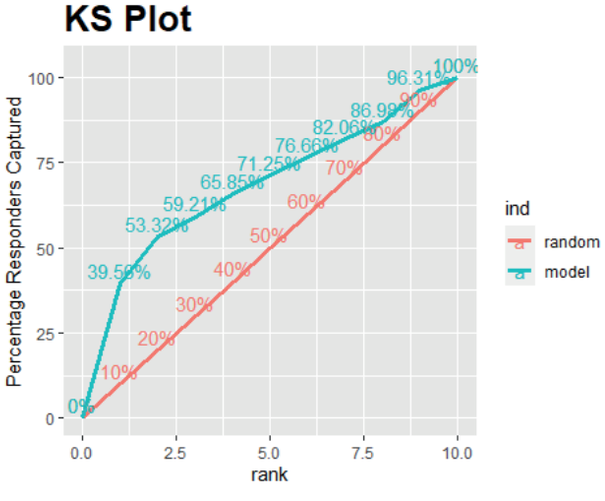

Рис. 3.9. KS Plot для моделі ідентифікації потенційного споживача на ринку електронної комерції

У табл. 1 можемо бачити інші оцінки помилок моделі та зробити висновок, що в цілому модель є адекватною.

Таблиця 1

Оцінки помилок моделі градієнтного бустінгу

Mean squared error	0.1315
Root mean squared error	0.3626
Relative squared error	0.8113
Mean abs error	0.1315
Mean abs deviation	6.575e-05
Relative abs error	0.4056

Висновки. Модель правильно ідентифікувала 1541 з 1752 клієнтів, які нічого не придбали на сайті, та 196 з 248 клієнтів, які здійснили покупку. Середня точність моделі дорівнює 88,1 %. Тобто ми справді підібрали правильну модель, яка добре ідентифікує клієнтів компанії і підходить для даного дослідження.

Дослідження поведінки споживачів може надати цінну інформацію для компанії. Компанія може ідентифікувати потенційних споживачів та зосередити свої сили саме на них, отримуючи вищу ефективність та більше продажів. Також ця інформація є цінною для маркетингового відділу та допомагає запустити цільові рекламні кампанії, знаючи портрет потенційного покупця. За допомогою градієнтного бустінгу можна не просто ідентифікувати споживачів, а також сегментувати їх на різні групи, якщо у компанії є достатня кількість потрібної інформації.

Література:

1. Understanding Online Consumer Behaviors for a Better Customer Journey. <https://www.shipbob.com/blog/online-consumer-behavior/ > (2021,November,18).

Understanding Online Consumer Behaviors for a Better Customer Journey. <https://www.shipbob.com/blog/online-consumer-behavior/ > (2021,November,18). [in English]

2. Johnson, R. & Zhang, T. (2014). Learning Nonlinear Functions Using Regularized Greedy Forest. IEEE Transactions on Pattern Analysis & Machine Intelligence, 36(5), 942-954. (2021,November,11)

Johnson, R. & Zhang, T. (2014). Learning Nonlinear Functions Using Regularized Greedy Forest. IEEE Transactions on Pattern Analysis & Machine Intelligence, 36(5), 942-954. (2021,November,11) [in English]

3. Ecommerce Dataset. Kaggle datasets. <https://www.kaggle.com/lissetteg/ecommerce-dataset>. (2021, November, 05)

Ecommerce Dataset. Kaggle datasets. <https://www.kaggle.com/lissetteg/ecommerce-dataset>. (2021,November,05). [in English]

4. Training and Test Sets: Splitting Data | Machine Learning Crash Course. <https://developers.google.com/machine-learning/crash-course/training-and-test-sets/splitting-data> (2021,November,15).

Training and Test Sets: Splitting Data | Machine Learning Crash Course. <https://developers.google.com/machine-learning/crash-course/training-and-test-sets/splitting-data> (2021,November,15). [in English]

5. Markham K. Simple guide to confusion matrix terminology. Data School. <https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/> (2021,November,09).

Markham K. Simple guide to confusion matrix terminology. Data School. <https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/>. (2021,November,09). [in English]

6. Kolmogorov-Smirnov Goodness of Fit Test. Statistics How To. <https://www.statisticshowto.com/kolmogorov-smirnov-test/>. (2021,November,13).

Kolmogorov-Smirnov Goodness of Fit Test. Statistics How To. <https://www.statisticshowto.com/kolmogorov-smirnov-test/>. (2021,November,13). [in English]